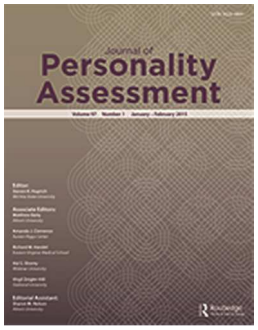
Viglione, Donald J; Giromini, Luciano; Landis, Patricia. The Development of the Inventory of Problems–29: A Brief Self-Administered Measure for Discriminating Bona Fide From Feigned Psychiatric and Cognitive Complaints. JOURNAL OF PERSONALITY ASSESSMENT. None pp: 1-11.
DOI: 10.1080/00223891.2016.1233882



The Development of the Inventory of Problems-29: A Brief Self-Administered Measure for Discriminating Bona Fide from Feigned Psychiatric and Cognitive Complaints

Journal:	*Journal of Personality Assessment*
Manuscript ID	JPA-2016-095.R1
Manuscript Type:	General Submission
Keywords:	Inventory of Problems, Feigning, Malingering < Content or Topic, Test Development < Content or Topic, Validity

DEVELOPMENT OF THE IOP-29

Abstract

This article describes the development of the Inventory of Problems-29 (IOP-29), a new, short, paper-and-pencil, self-administered measure of feigned mental and cognitive disorders. Four clinical comparison, simulation studies were conducted. Study 1 (n = 451) selected the items and produced an index of potential feigning. Study 2 (n = 331) scaled such index to produce a probability score, and examined its psychometric properties. Study 3 tested the generalizability of Study 2's findings with two additional samples (ns = 128 and 90). Results supported the utility of the IOP-29 for discriminating bona fide from feigned psychiatric and cognitive complaints. Validity was demonstrated with mild traumatic brain injury, psychosis, PTSD, and depression. Within the independent samples of studies 2 and 3, the brief IOP-29 performed similarly to the MMPI-2 and PAI, and perhaps better than the TOMM. Classifications within these samples with base rates of .5 produced sensitivity, specificity, positive predictive power, and negative predictive power statistics of about .80. Further research is needed testing the IOP-29 in ecologically valid field studies.

Keywords: Inventory of Problems, feigning, malingering, test development, validity

The Development of the Inventory of Problems-29: A Brief Self-Administered Measure for Discriminating Bona Fide from Feigned Psychiatric and Cognitive Complaints

Malingering of mental and cognitive disorders costs millions of dollars each year (Chafetz & Underhill, 2013). Undetected cases of malingering are costly to taxpayers, as funds, for example, are allocated to provide malingering criminals with housing and unnecessary psychiatric treatment in hospitals rather than prisons. Insurance companies and businesses fund treatment for malingered neuropsychological and psychological complaints, as well as pay awards in malingered workman's compensation cases. To some degree, malingering compromises the efficacy of the mental health system, as practitioners may not be able to provide accurate diagnoses, recommendations, or effective treatment to those malingering psychological complaints. Also, true patients may be viewed with a suspicious eye and have more difficulty accessing treatment and compensation.

Reported rates of malingering vary greatly, because malingering involves both "situation-specific and issue specific" (Rogers & Salekin, 1998, p.148), there is a strong motivation to avoid detection, and different detection methods are used. In a review of literature, Rogers (1997) cited two large surveys estimating malingering prevalence rates to range between 15.7% and 17.4% among forensic settings. Nevertheless, these rates are in fact highly variable in forensic practice, with a standard deviation of 14.4% (Rogers, Salekin, Sewell, Goldstein, & Leonard, 1998). More recently, Larrabee (2003) grossly estimated the average rate of neuropsychological malingering from research publications to be 41%, although estimates in the individual studies differed by as much as 50%.

Given the importance of the phenomenon, it is not surprising that research on the detection of malingering in the form of feigning on psychological tests, has dramatically increased during the past few decades, and a number of instruments and scales to detect

feigning[1] of mental disorders, cognitive impairment, and medical complaints have been proposed (Rogers, 2008). Multiscale personality inventories such as the Minnesota Multiphasic Personality Inventory (MMPI-2; Green, 1991; MMPI-RF; Ben-Porath & Tellegen, 2008) and Personality Assessment Inventory (PAI, Morey, 1996) include validity indices aimed at detecting atypical response styles and exaggeration. Some interview-based and multi-method assessment procedures specifically focus on detecting feigning of mental disorders also have been described. Among them, Rogers and his coworkers (Rogers, Gillis, Dickens, & Bagby, 1991; Rogers, Gillis, Bagby, & Monteiro, 1991; Rogers, Sewell, & Gillard, 2010) have developed the Structured Interview of Reported Symptoms (SIRS), an instrument to detect feigning of mental disorders that has received widespread acceptance in the forensic community (Archer, Buffington-Vollum, Stredney, & Handel, 2006; Lally, 2003).

Because these instruments require either extended administration time or trained examiners, several brief and easy to use instruments to detect feigning of mental disorders have also been introduced. For example, Miller (2001) developed the Miller-Forensic Assessment of Symptoms Test (M-FAST), a 25-item structured interview for feigned psychopathology which has demonstrated good scale characteristics. As another example, Beaber and his associates (Beaber, Marston, Michelli, & Millis, 1985) developed the M Test, a short, self-administered, self-report measure, designed to expose false psychotic complaints. Along the same lines, Smith and Burger (1997) developed the Structured Inventory of Malingered Symptomatology (SIMS), a 75-item self-report tool aimed to detect commonly feigned conditions such as psychosis, neurological impairment, and affective disorders.

[1] Consistent with Rogers and Bender (2013), here we refer to *malingering* to indicate the "deliberate fabrication or gross exaggeration of psychological or physical symptoms of the fulfillment of an external goal" and *feigning* to indicate the "deliberate fabrication or gross exaggeration of psychological or physical symptoms (Rogers & Vitacco, 2002) without any assumptions about its goals." (p. 518)

In addition, a number of brief measures also have been developed with the purpose to detect feigned cognitive impairment. These instruments, sometimes referred to as "tests of effort" or "performance validity tests," generally include cognitive tasks that may appear difficult at a first sight, but are in fact relatively simple. Thus, by exerting inadequate or suboptimal effort, feigners often perform more poorly than individuals with genuine cognitive impairment. Examples of these instruments include the Test of Memory Malingering (TOMM; Tombaugh, 1996), the Victoria Symptom Validity Test (VSVT; Slick, Hopp, Strauss, & Thompson, 2005), and the Word Memory Test (WMT; Green, Allen, & Astner, 1996).

All of these instruments vary in their merits, foci, and limitations. For example, some, such as the TOMM or the WMT, only focus on certain, specific cognitive abilities (e.g., recognition memory) and are less likely to detect feigned affective disorders or psychosis. In contrast, others, such as the M Test, are designed to detect a specific, single diagnostic target, such as feigned psychosis, but may not detect other feigned cognitive or affective problems. Both cases are somewhat problematic, in that individuals attempting to feign mental or cognitive disorders often present with a multitude of symptoms and problems within a specific life predicament or context rather than limiting their efforts to a single, specific diagnostic target. Furthermore, some of these screening tools have mostly been studied using simulators[2] and honest, non-patient controls, while for the sake of generalizability to real life situations it would be optimal if simulators were contrasted to and matched with genuine patients. Also, many focus on only a single detection strategy, possibly under the influence of internal consistency concerns, so that their items do not vary systematically to provide comprehensive coverage of tactics adopted by malingerers (Rogers, 2008). Perhaps more importantly, all suffer from some errors in prediction, shrinkage from developmental research

[2] The word *simulators* is used here to characterize research participants in simulation studies feigning any psychiatric disorders or cognitive impairments; in contrast, the word *malingerers* is used to describe individuals presenting disorders or impairments in real life.

to validation and field research, and uncertainty regarding the cutoff scores across situations and studies, the last of which is a routine problem in decision-making in practice.

In an attempt to overcome some of these limitations and to help examiners decision-making in the field, we developed the Inventory of Problems-29 (IOP-29), a short, paper-and-pencil, self-administered measure to detect feigning of mental or cognitive disorders. With this first, IOP-29 study, we intended to provide evidence that such measure can detect feigning of both psychiatric and cognitive disorders within simulation studies.

Overview

Four guiding principles shaped the development the IOP-29. First, we conceptualized *malingering* as a person-situation interaction with a developmental course over a period of time, reinforced through powerful incentives (Rogers, 1988). Secondly, since malingerers adopt different strategies and situations induce various approaches from individuals (Rogers & Bender, 2013; Rogers, Harrell, & Liff, 1993), we concluded that a short, but comprehensive feigning measure would necessitate multiple detection strategies. Third, an implication of these principles is that test and person interact in a particular context or predicament. Within this interaction, malingering behaviors may ebb and flow as a function of environmental stimuli. Potentially, a test can induce and measure feigning behaviors that otherwise might not be available. Finally, consistent with Arbisi and Ben-Porath (1995), we focused on incremental validity in developing items and selecting items and strategies. Thus, in the test development we retained items that incremented over the easier to develop keyed true, atypical or extreme complaint items. Such descriptions dominate many self-report malingering tests.

From the sentiment in the field about the advantages of interviewing techniques and the research supporting the validity of complex interviewing techniques (e.g., Rogers, 2008), we concluded that we had to create a different type of objective test. Such a test might

incorporate "proven" techniques from interviews. Thus, in our test development studies leading up to the IOP-29, we investigated a number of different item structures and organizational methods. First, we included a variety of cognitive problems along with the more conventional test item declarative statements. Problem-solving included memory, calculation, and reasoning items (e.g., "Reading is to a Book, as Listening is to a Song"). Secondly, we added the response choice of "doesn't make sense" to the typical true-false response selections. We hypothesized that this three response choice structure would allow more precise identification of feigned versus true psychopathology. In addition, it potentially allowed us to address the simulators' use of extreme or nonsensical complaints more easily, so that honest test takers could answer "doesn't make sense" to items assuming a malingering response set (e.g., "It never stops torturing me"). Alternatively, feigned cognitive deficiency and resistance to the evaluation might emerge as frequent endorsement of "doesn't make sense" (Rogers, 1997). Furthermore, responses to problem-solving items were not restricted to this "true-false-doesn't make sense" response choice, allowing us to incorporate approximate answers or near-miss phenomena (e.g., "Answer this problem: 165 – 121 = ?"). We also attempted to address "test-behaviors" that we might induce (and simulators might self-report) during the test itself. For example, we speculated that simulators presenting false disorders might note that they are shaking uncontrollably or that they might report concerns about the test not surveying all of their problems.

An extensive series of studies and three prior versions of the instrument (Viglione & Landis, 1994) contributed to the development of the IOP-29. This long developmental period adapted concepts, techniques, and statistical procedures derived from the literature and empirical experience with the versions of the scale. Our focus was on ecological validity and utility by focusing on helping examiners in the field in forming an opinion about whether the

presenting complaint is bona fide or feigned[3]. In each stage of our research, strategies, techniques, items, and item combinations were revised, trimmed, and enhanced based on the empirical findings. We concerned ourselves with discriminating simulators from patients, with little emphasis on honest, non-patient controls, as they are very easy to identify. To maximize generalizability of our findings, our research studies utilized adult, non-college, community samples and included instructions not to overdo feigning roles.

Initially, based on our review of the literature and experience, we developed 27 potential feigning strategies and 245 items (Viglione & Landis, 1994). These strategies included concepts such as: (a) endorsement of atypical symptoms, that do not occur or are very rare even with significantly disturbed individuals; (b) failure of easy cognitive problems, which bona fide patients and non-patient controls almost always answer correctly; (c) externalization or minimization of one's own responsibility concerning his or her own psychological condition or situation; (d) criticism or nervousness of or problematic reactions to the evaluation context, the test itself, or the testing condition; and (e) refusal to admit qualified positive attributes or experiences, which bona fide patients and non-patient controls almost always endorse. Thus, using Rogers and Bender's (2013) words, we combined both *unlikely* and *amplified* mental disorder detection strategies with both *excessive impairment* and *unlikely presentations* strategies to feign cognitive impairment within a single instrument to maximize prediction. Moreover, many items were worded in ways to capture a variety of disorders, and each item corresponded to one or more feigning detection strategies to serve our goal of using the test to address a wide variety of disorders. This initial research was an attempt to determine which of these detection strategies might work in identifying feigning. It included 226 simulators, a few suspected, malingerers in vivo, and patients.

[3] In practice, of course, this decision is complicated by both multiple, related response styles, and the fact that motivation, i.e. malingering, cannot be deduced from test findings. Also, findings from individual tests are often indeterminate regarding the likelihood of feigning (Rogers & Bender, 2013).

Based on the findings, we retained those strategies, items, and item and response structures that discriminated well, we eliminated those that did not, and modified others. We added items for empirically supported detection strategies, modified items so that they accessed multiple strategies, and added items according to the emerging research in the field. Thus, we created a second, developmental version of the instrument (IOP-Developmental Version-2; IOP-DV2), which consisted of 162 items intermixing a minority of cognitive items with self-report items. Noteworthy, the great majority of these 162 items had already demonstrated validity in identifying simulators, with a smaller number of new items being based on generalizations from strategies and item types that had demonstrated validity in our first pilot research.

The Current Studies

The main goal of the current group of studies was to develop, scale, and cross-validate the IOP-29, a brief, self-administered measure of feigning. Its main application is to serve as a decision-making tool to discriminate bona fide patients from feigners. We present three test development, simulation studies with various independent samples. The first study was conducted to select a small subset of items which together best discriminate feigning from bona fide disorders. We thus selected the 29 items of the IOP-29. The second study was conducted to scale the IOP-29 and produce a feigning probability score. The third study cross-validated the probability score obtained from Study 2. All studies included clinical comparison samples.

Study 1: Scale Construction

Method

To select the best subset of the 162 items, which as a group would optimize the validity of the scale, we retrieved archival data from eight studies. Each study used specific procedures and measures, and investigated different samples with different purposes.

Nevertheless, all included data on the IOP-DV2 and followed relatively similar procedures in terms of inclusion and exclusion criteria, diagnostic status verification, and simulation task presentation. For the purpose of this study, we only extracted from each study the IOP-DV2 data and only included patients' and simulators' data.

Sampling procedures and participants. The sample utilized for Study 1 combined data (n = 451) from six dissertation studies (Green, 1999; Jansak, 1996; Landis, 1996; McDougall, 1996; Mellin, 1996; Schaich, 2000) and two unpublished studies, for a total of 160 bona fide patients and 291 simulators. Most of the data, 76%, were collected with a computerized version of the IOP-DV2 (e.g., Landis, 1996) with 24% being collected with a paper-and-pencil version (e.g., Jansak, 1996). In most of the cases, simulators and to a lesser degree patients were recruited from the general community through newspapers, advertisements, and flyers. Most of the patients (e.g., Mellin, 1996; Schaich, 2000) were instead recruited through inpatient hospitals, outpatient clinics, day treatment centers, or criminal investigation units. In most cases (i.e., Green, 1999; Jansak, 1996; Landis, 1996; McDougall, 1996; Mellin, 1996; Schaich, 2000) participants were compensated for their participation with a small amount of money.

All diagnoses of the patient group were verified either by the chief evaluator of the referring site, or by the hospital records, or through some consolidated assessment instruments such as the Structured Clinical Interview for the DSM-IV (SCID-IV; First, Spitzer, Gibbon, & Williams, 1995), or the PTSD Interview (Watson, Juba, Manifold, Kucala, & Anderson, 1991). In almost all cases, criteria for inclusion in the simulator group included no evidence of severe psychopathologies or active substance abuse disorder. For some of the studies, additional inclusion criteria for the simulator group also included no evidence of a specific "target" diagnosis. Landis' (1996) study, for example, aimed at contrasting bona fide PTSD patients to PTSD simulators, and also excluded potential

simulators if they exceeded the PTSD severity criteria. In all cases, simulators were given a brief scenario with descriptions of symptoms to provide a life-like motivation to malinger. For example, simulators of depression were asked to imagine that they had had an accident at work, that such accident was caused by their boss' negligence, that their boss decided to unreasonably limit their recovery time (thereby cutting their disability payments), and that their only chance to remain on disability would be to fake depression. To limit extreme, unrealistic portrayals of disorders by simulators and to minimize an artificially large effect size that might limit generalization to real life situations, simulators were warned that if they presented their symptoms too dramatically, their performance would not be believable (Rogers & Bender, 2013 Viglione, Wright, Dizon, Moynihan, DuPuis, & Pizitz, ,2001). Finally, in most of the cases simulators were also told that the most "successful feigners," that is those who were not detected, would be put in a lottery and could win additional financial rewards.

Most of the patient group had a diagnosis of depression (≈ 46%), a diagnosis in the schizophrenia spectrum (≈ 27%), or a diagnosis of PTSD (≈ 18%). However, a small proportion (≈ 9%) also included other diagnoses (e.g., ≈ 7% had a bipolar disorder). Similarly, most of the simulator group were non-patients who faked depression (≈ 25%), schizophrenia (≈ 39%), or PTSD (≈ 20%). The rest of the simulator group consisted of incarcerated inmates simulating incompetence to stand trial (≈ 10%) or non-patient simulators faking other disorders (≈ 6%).

Individual demographic information was available, albeit with some missing values, for 282 out of 451 cases. By supplementing these data with group and sample demographic descriptions reported by Green (1999), Jansak (1996), Landis (1996), McDougall (1996), Mellin (1996), and Schaich (2000), in their doctoral dissertations, we were able to produce a relatively accurate characterization of the entire sample: Within the patient sample, ages

ranged from 20 to 79 years old, with a mean age of 38.8 (*SD* = 11.5); within the simulator sample ages ranged 16 to 80 years old, with a mean age of 33.8 (*SD* = 10.7). Although such a difference is statistically significant, $t_{(449)} = 4.6$, $p < .01$, $d = .45$, it consists of merely five years, which in adulthood is likely to not affect the IOP-DV2 data. Gender did not differ, *Phi* = -.01, $p = .89$, with about 52% males and about 48% females in the patient sample and about 51% male and 49% female among simulators. Within the patient sample, about 26%, were non-Caucasians, with about 12% of African-Americans, about 8% of Hispanics, about 3% of Asian-Americans, and about 3% of other ethnicities. Within the simulator sample, a similar proportion, i.e., about 24%, of non-Caucasians was observed, with about 12% of Hispanics, about 4% of Asian-Americans, about 3% of African-Americans, and about 5% of other ethnicities. The percentage of Caucasians vs. non-Caucasian within the patient and simulator samples, thus, did not statistically differ, *Phi* = -.01, $p = .77$. Within the patient sample about 53% had at least attended some college, whereas such percentage within the simulator sample was about 75%. Thus, the simulator group was more educated than the patient group, *Phi* = -.23, $p = <.01$. Finally, about 14% of the patient sample was married, while about 23% of the simulator sample was married. This difference, again, is statistically significant, *Phi* = -.10, $p = .03$, although the small effect size was small.

Item selection. Aiming at developing a brief, paper-and-pencil self-report measure, a sub-set of the initial 162 IOP-DV2 items was selected. To do so, we followed a three-step procedure. First, we excluded all items that were not suitable for a paper and pencil self-administration. For example, we excluded some memory items in that they required images to be shown on a computer screen. This initial screening yielded 126 items to be further examined. Secondly, we calculated the Phi correlations between each item (1 = response in the key direction; 0 = response not in the key direction) and group (dummy code 1 = simulator; 0 = bona fide patient). Item/response combinations with significant Phi values ($p \leq$

.05) and in the expected direction were retained for the third step of our item selection procedure. Thus, a given keyed true item was retained only if the response "true" (1 = true; 0 = not true) produced a positive and significant Phi correlation with group, and, similarly, a given keyed false item was retained only if the response "false" (1 = false; 0 = not false) produced a positive and significant Phi correlation with group. Although the item selection process focused on the true vs. false options, the additional "doesn't make sense" choice option effects this process. Potentially, this third response option allows for a more refined and less ambiguous understanding of the meaning of the true and false response selection by eliminating its forced choice character and reducing error.

We found 114 significant item/response Phi correlations among the 126 items, a testament to the validity of the item development and validity of our procedures to this point. We implemented a relatively complex algorithm to extract a small subset of items that together would maximize the IOP-29 correlation with group while minimizing redundancy among the items. The first item selected is the item/response combination with the highest Phi value, i.e., the best predictor. Subsequent items were those with the best incremental validity. To select those items one-by-one, we partialled out previously selected item(s) from the remaining item/response-group correlations, and selected the item with the highest partial Phi correlation. This procedure was stopped after selecting 29 items, when the highest remaining partial Phi correlation approached zero.

Item weights. With these 29 items, an initial feigning index (Single Weight Score) was produced by summing them with one point for each item/response combination. We then selected all item/response combinations from these 29 items that demonstrated incremental validity – operationally defined by a significant Phi correlation in the expected direction after partialling out the Single Weight Score. This procedure resulted in multiple weighting for the subset of items with strong validity and involved the "doesn't make sense" response more

directly in the scaling. For example, a given item keyed false may have a +1 weighting for "false", a 0 weighting for "doesn't make sense" and a -1 weighting for "true." The resulting index was named Multiple Weight Score.

Results

The point bi-serial correlations between group (dummy code 1 = simulator; 0 = bona fide patient) and the two feigning indexes under investigation were respectively .687, $p <$ .001, for the Single Weight Score, and .714, $p < .001$, for the Multiple Weight Score. The difference between these two correlations was significant ($z = 2.7$ $p < .001$), using procedures described by Meng, Rosenthal, and Rubin (1992) for comparing correlations obtained from the same sample that share a common variable. The receiver operator characteristic curve (AUC) for the Single Weight Score was .913 (*SE* = .014), the AUC for the Multiple Weight Score was .932 (*SE* = .012). Cohen's *d* values were 1.97 for the Single Weight Score and 2.13 for the Multiple Weight Score.

Study 2: Probability Scaling

Study 2 utilized the Multiple Weight Score derived from Study 1 to produce and to cross-validate a False Disorder Probability Score. That is, a second, independent sample was used to generate the transformations of raw scores to the scale used for the IOP-29, a probability of feigning score. The independence of the samples in Study 1 and 2 prevents distortions of the normative scaling. Similar to Study 1, data were retrieved from a number of studies that used a variety of procedures and measures. Again, all included studies followed relatively similar inclusion and exclusion criteria, diagnostic status verification, and simulation task presentation. Also consistent with Study 1, for the purpose of Study 2 only the IOP data were utilized and data from non-patient, honest participants was excluded. Half of this dataset was randomly selected to produce the False Disorder Probability Score, and the other half was used to cross-validate this score.

DEVELOPMENT OF THE IOP-29

Development of the False Disorder Probability Score

Sampling procedures and participants. Subsequent to creating the 162 items of the IOP-DV2, we tested some additional items and developed a third, developmental version of the instrument (IOP-Developmental Version-3; IOP-DV3), a 181 item, computer-administered feigning measure. This version of the IOP includes all the 29 items that form the IOP-29[4]. Thus, to produce the IOP-29 False Disorder Probability Score, the current study investigated data from four simulation studies conducted with the IOP-DV3. Each study addressed one of the four disorder categories for which the IOP-DV3 was designed, i.e., neuropsychological deficit (Pizitz, 2001), schizophrenia and psychosis (O'Brien, 2004), Post-Traumatic Stress Disorder (Connell, 2004), and depression (Abramsky, 2005). Moreover, some combination of these four likely addresses a great majority of the malingering forensic clinical practice. All four studies used a simulation design contrasting a clinical comparison group of approximately 40 bona fide patients to approximately 40 individuals feigning the disorder. Participation was on a voluntary basis with prior consent obtained before participating, and in three of the four a mall monetary compensation was given to all participants. Procedures used for these four studies are very similar to those described above for Study 1 (for additional details, see Abramsky, 2005; Connell, 2004; O'Brien, 2004; and Pizitz, 2001). As reported in Table 1, patients and simulators in this large (n = 333), combined sample did not differ from each other in terms of gender, age, ethnicity, education, and marital status.

As noted above, to produce our feigning probability score, and cross-validate it with an independent sample, we split this N = 331, Study 2 sample. Half of each the patient and simulator samples within each study were randomly assigned to the developmental sample or validation sample. Thus, the developmental sample included 168 participants, 84 of which

[4] Accordingly, both the DV2 and DV3 versions of the IOP contain the IOP29 items. Both also included other items being evaluated for a longer more complex test of feigning.

were bona fide patients and 84 simulators. Ultimately, missing data reduced this number to 166 participants, 84 patients and 82 simulators, with approximately 20 patients and simulators in the four diagnostic groups: mild traumatic brain injury (MTBI) as an example of cognitive or neuropsychological disorders, schizophrenic spectrum-psychosis patients, PTSD, and depression.

Results. The Multiple Weight Score from Study 1 was investigated with the developmental sample of Study 2. The point bi-serial correlation between group (dummy code) and the Multiple Weight Score was .625, $p < .001$. The difference in the score of simulators and patients corresponded to a Cohen's *d* of 1.59. The AUC was .868 (*SE* = .029).

To produce a feigning probability score, the Multiple Weight Score was then used as predictor in logistic regression with group as the outcome. The model was significant, $\chi^2_{(1)} = 78.51$, $p < .001$, with a Nagelkerke R^2 of .502. The *exp(B)*, or odds ratio of the Multiple Weight Score was 258,483. Using a cut-off of .50, which nearly matches the feigning base rate in this sample (.494), the resulting, False Disorder Probability Score correctly classified 81.3% of the cases, with 14 patients erroneously classified as simulators (i.e., 8.4% of the total, and 17.7% of those predicted to be malingering), and 17 simulators erroneously classified as patients (i.e., 10.2% of the total, and 19.5% of those predicted to be not malingering). Sensitivity is .79, specificity .83. Using the base rate observed in our sample (i.e., .494), positive predictive power (PPP) is .82, negative predictive power (NPP) is .80, and overall correct classification (OCC) is .81. Consistent with findings obtained for the Multiple Weight Score, the AUC was .868 (*SE* = .029) also for the False Disorder Probability Score.

Cross-Validation of the False Disorder Probability Score

Next, we tested the validity of our feigning indexes with the independent Study 2 validation sample consisting of 165 participants with 82 bona fide patients and 83 simulators.

Results. The point bi-serial correlations between group (dummy code) and the Multiple Weight Score was .625, $p < .001$; the difference in the score of simulators and patients yielded a Cohen's *d* of 1.67; the AUC was .867 (*SE* = .029). These findings are virtually identical to those obtained from the developmental sample, so that no shrinkage was shown, thus lending credence to our item development and item weighting procedures. With a cut-off of .50, the False Disorder Probability Score correctly classified around 80% of the cases, with 17 false positive classifications (i.e., 10.3% of the total, and 20.2% of the positives), and 16 false negative classifications (i.e., 9.7% of the total, and 19.8% of the negatives).

Sensitivity, specificity, PPP, NPP, and overall correct classification (OCC) statistics were calculated. Due to the PPP, NPP, and OCC statistics being highly dependent on the prevalence or base rate of the condition being tested (Meehl & Rosen, 1955), in addition to the approximately .50 base rate[5] of our sample, base rates of .25, .10, and .05 were also explored by adjusting the PPP, NPP, and OCC formulas accordingly (Streiner, 2003). These additional base rates were selected in that they may present a priori base rates or prevalence in a variety of clinical, forensic, disability, benefit, or compensation settings. In addition, we also surveyed probability score cutoffs, because they influence positive and negative predictive power, and false positive and false negative errors are associated with different cost-benefit ratios. Results, reported in Table 2, indicate that when a cut-off of .50 is adopted, about 80% of the simulators are detected (sensitivity = .81), and about 80% of the patients are correctly labeled as non-simulators (specificity = .79). Also, the OCC is maximized when the cut-offs under examination reflect the base rates. For example, when a base rate of .50 is considered, the OCC is maximized (OCC = .80) if the cut-off of the False Disorder

[5] The base rate for malingering in Study 3 is .503.

Probability Score is .50; similarly, when a base rate of .10 is considered, the OCC is maximized (OCC = .92) for a False Disorder Probability Score of .90.

Comparative validity. For about half of the samples utilized for Study 2, we also had available data on frequently used feigning scales to address the comparative validity of the IOP-29 against the MMPI-2 and TOMM. Specifically, O'Brien's (2004) psychotic sub-sample had MMPI-2 data from 43 patients and 45 simulators, and Abramsky's (2005) depression sub-sample included TOMM data from 43 patients and 42 simulators. To maximize the stability and reliability of the statistics, comparative validity analyses were conducted using all the available participants (i.e., 88 for the psychotic sub-sample, and 85 for the depression sub-sample). Because about half of these data were also included in the developmental sample of Study 2 and used to produce the final scaling of the False Disorder Probability Score, however, in addition to False Disorder Probability Score also the Multiple Weight Score was examined. This latter score is completely independent of the scaling in the second study, so could not possibly provide any over-estimation of the effect size.

As for the MMPI-2, consistent with O'Brien's (2004) work, the F, Fp, and Ds-r2 scales were selected for these analyses, as they are particularly useful in detecting feigning of schizophrenia or psychosis (e.g., Elhai, Gold, & Frueh, 2000; Rogers, Sewell, & Ustad, 1995; Viglione et al., 2001). As for the TOMM, consistent with Abramsky's (2005) work as well as with previous research (e.g., Rees, Tombaugh, Gansler, & Moczynski, 1998; Tombough, 1996), both the number of correct responses on the first trial (TOMM-1), second trial (TOMM-2), and entire test (TOMM-tot) were inspected.

The correlations calculated between each measure and group are presented in Table 3, along with the respective AUC values. These data indicate that the IOP-29 performs similarly to the selected MMPI-2 scales, and perhaps better than the TOMM. Also noteworthy, the

IOP-29 performed similarly well for all of the four diagnostic groups taken into consideration for this study, i.e., MTBI, Schizophrenia/Psychosis, PTSD, and Depression.[6]

Study 3: Generalizability of the Findings

The analyses performed with the validation sample of Study 2 provide initial cross-validation of the IOP-29 False Disorder Probability Score. However, because they utilized the randomly excluded half of the subsamples used to develop the feigning probability score, Study 3 was designed to further validate the IOP-29 with independent samples.

Method

Sampling procedures and participants. Study 3 included two additional, independent samples from two dissertations. The first was a forensic sample of 128 volunteer participants on community-based county and federal probation (McCullaugh, 2011); half responded genuinely, half simulated mental health symptoms. The second sample included 90 participants: half were patients with Schizophrenia or Schizoaffective Disorder and half simulators (Wood, 2008). For both samples, the IOP-DV3 and Personality Assessment Inventory (PAI; Morey, 1991, 1996, 2007) were administered. To avoid extreme, unbelievable performances, participants were warned not to "overdo" their presentations, otherwise their performance would not be believable (Viglione et al, 2001).

McCullaugh's (2011) forensic sample. Participants (n = 128) were volunteer offenders on community-based county and federal probation. Twenty-four were recruited from a local, private forensic psychologist's practice; all were being treated using a mixture of manual-based cognitive behavioral therapy (CBT) treatment for sexual offenders and substance abuse treatment. Confirmation that all these individuals were on probation was provided by the treating psychologist. The remaining 104 participants were recruited via

[6] As reported in Table 3, AUC was .845 (*SE* = .042) for the psychotic sub-sample and .898 (*SE* = .037) for the depression sub-sample. For the PTSD sub-sample AUC was .903 (*SE* = .037); for the MTBI sub-sample AUC was .840 (*SE* = .048).

volunteer postings on the internet. To verify that they also were on probation, the San Diego Country Court Index website was used.

Half of the sample (randomly selected) was asked to respond genuinely, the other half was asked to simulate mental health symptoms. Inclusion criteria required that participants were not prescribed anti-psychotic medication, did not experience psychotic symptoms, and did not report substance use within the last 30 days. Approximately 80% were men, the mean age was about 36 (SD approximately 10), and about 70% were Caucasian. No significant differences between the two groups were found for age, gender, ethnicity, education, marital or employment status. Following completion of the procedures, all participants were debriefed about the aims of the study, provided with $20.00 incentive for participation, and thanked for participating. Two additional raffle winners were drawn after study completion and provided with $100.00 cash prizes.

Consistent with Rogers and Gillard's (2011) recommendations, simulators were given a vignette to facilitate feigning. Based on details of an actual case of suspected malingerer from another study (Weber, 2008), these vignettes were rich in contextual and historical detail. Raters judged these vignettes as a group to include diffuse mental health symptoms, including post-traumatic stress, neuropsychological impairment, and depressive symptoms, but not schizophrenic or psychotic ones. For example, one of the vignettes asked the individual to imagine having suffered an injury while working at a horse stable. The injury resulted in being laid off without receiving disability income. Faking mental health symptoms would be the way to receive compensation benefits for his or her disability. All simulators were asked to carefully read these vignettes, and respond to the measures as the individual depicted, reflecting history and symptom presentation.

Woods' (2008) psychotic sample. The second sample included in Study 4 is a psychotic sample, derived from a dissertation study (Wood, 2008) conducted with the IOP-

DV3. Patients (n = 45) were recruited from local day treatment programs, and all had a diagnosis of Schizophrenia or Schizoaffective disorder, as confirmed by their treating psychiatrists. Simulators (n = 45) were matched with patients on gender and level of education: In each group there were 34 men and 11 women, 36 had no college education and 11 did. The mean age within patients was 40 years (SD = 10), while the simulators were younger, with a mean age of 28 years (SD = 10). In both groups of this diverse sample, about one third of the participants identified themselves as Caucasian, about one third as African-American, and about one third as Latino. Approximately 70%, in each group, were single. All patients, but none of the simulators, had a history of psychiatric hospitalization and were currently taking psychotropic medication. Prior to taking the IOP, simulators were given ten minutes to review the instructions, which included a description of the disorder. In short, they were asked to complete the psychological testing as if they were trying to appear psychotic in a given real-life scenario (e.g., they were asked to pretend they had been accused of committing a crime, and that the only way to avoid incarceration would be to appear psychotic). Furthermore, simulators were also told that the best fakers would receive an additional $50.00, in addition to the $20.00 compensation given to all participants.

Measures. Both McCullaugh' (2011) and Wood' (2008) also used the PAI, a 334-item, self-report, personality inventory widely used in research on malingering. Positive findings for the PAI capacity to identify malingering has been reported consistently in the literature (e.g., Archer et al., 2006; Blanchard et al., 2003; Sellbom, Bagby, & Rogers, 2008). According to a recent meta-analysis (Hawes & Bocaccini, 2009), the Negative Impression Management (NIM; Morey, 1996) and the Malingering Index (MAL; Morey, 1996) have demonstrated the best overall classification rates for malingering. Specifically, a NIM cut score of T ≥ 81 resulted in an average hit rate of .79 (SE = .09), sensitivity of .73 (SE = .13), and specificity of .83 (SE = .07); a cut score of MAL ≥ 3 yielded an average hit rate of .71

(*SE* = .05), sensitivity of .58 (*SE* = .08), and specificity of .86 (*SE* = .04). Additionally, the Rogers Discriminant Function (RDF; Rogers, Sewell, Morey, & Ustad, 1996) also has demonstrated some clinical utility in applied settings, even though it does not strongly correlate with the SIRS (Edens, Poythress, & Watkins-Clay, 2007). For the RDF, Morey (2007) suggested that raw scores above zero are indicative of risk of malingering.

Results

The point bi-serial correlation between the IOP-29 False Disorder Probability Score and group is comparable to that obtained with the PAI scales, and AUC analyses similar to for the PAI and IOP-29 (Table 4). Mean IOP-29 differences for simulators and patients yields a Cohen's *d* of 2.66 in the forensic sample, and a *d* of 1.95 in the psychotic sample. For the PAI NIM scale (T score), the Cohen's *d* values are 2.85 within the forensic sample, and 1.69 within the psychotic sample. For the MAL (raw score), Cohen's *d* was 1.36 within the forensic sample, and 1.78 within the psychotic sample. For the RDF, Cohen's *d* was 1.82 within the forensic sample, and 1.75 within the psychotic sample.

Using a cut score of .50 for the IOP-29, and cut scores recommended by Hawes and Bocaccini (2009) for the NIM and MAL, and by Morey (1996) for the RDF scales of the PAI, we then calculated diagnostic efficiency statistics for both the forensic and the psychotic samples. Results, reported in Table 5, confirm similar validity findings for both the PAI and the IOP-29, with the IOP-29 False Disorder Probability Score producing OCC rates equal to or greater than .80.

General Discussion

We presented development and psychometric foundation research for a new, brief, self-administered measure of feigning of a variety of disorders, the IOP-29. Study 1 selected the items to be retained from a pool of items of which most had already been validated as a measuring feigning. In addition, it produced a feigning index, the Multiple Weight Score.

Study 2 scaled this index and produced a probability score, the False Disorder Probability Score. Finally, Study 3 cross-validated the False Disorder Probability Score with two additional independent samples. What emerged from this series of studies is a 29-item scale to measure feigning of psychiatric and cognitive or neuropsychological disorders. It includes 26 self-report items and 3 cognitive items. We found a way to capture in vivo behavior that often is only accessed in interviews: Some of the most effective items are "test behavior" items, which refer to reactions to the test itself. Some items are worded in vague ambiguous ways with "it" or "problems" so as to capture malingering response style sets and a broad array or feigned difficulties. For self-report items, we included a third response alternative, namely "doesn't make sense," in addition to the standard "true" and "false," because we found it increased validity. Nonetheless, our investigation only focused on clinical comparison and simulation studies, therefore further research testing the IOP-29 in ecologically valid field studies is needed.

Taken together, the results of our studies lend initial support to the validity of the IOP-29 in discriminating bona fide patients from feigners. Within a number of independent, validation samples with base rates of about .50, and an a priori, statistically-determined cut score of .50, the IOP-29 correctly classified about 80% of the cases, with sensitivity, specificity, PPP, and NPP being around .80. Noteworthy, the samples under investigation came from a total of 14 independent studies, 12 doctoral dissertations and 2 unpublished studies, and were diverse in terms of age, sex, ethnicity, marital status, and educational background. The 29 items are an optimal combination selected from approximately 250 items with independent re-validations. Individual items were cross-validated at multiple stages in the test development process. Shrinkage in the later stages of development was minimal. In fact, the AUC of the False Disorder Probability Score was .87 both in the developmental and validation samples of Study 2 (SE's ≈ .03), and reached .94 (SE = .02) and .92 (SE = .03) in

the forensic and psychotic samples of Study 3. Thus, it is reasonable to anticipate that the IOP-29 will generalize well and perform similarly in different cultural contexts and future clinical comparison, simulation studies.

Most of the available feigning scales adopt classic test theory and normative scaling approaches and produce, for example, t-score comparisons between simulators (or suspected malingerers) and non-patient or patient groups. As a result, selecting optimal threshold scores for real life situations might be challenging, at times. The IOP-29 provides examiners with a fully cross-validated, easy to use probability score: The higher the False Disorder Probability Score, the higher the probability that the tested individual is trying to feign a psychological, psychiatric or neurological disorder. The interpretation of this score is simple and straightforward, and should contribute to the decision at hand, i.e., whether the presenting complaint is bona fide or feigned. However, users should keep in mind that while a cut-off score of .50 is likely to work best when the a priori probability is .50, other cut-off scores may be more appropriate for different a priori probabilities. For example, if in a given environment the expected prevalence rate of malingering is about .10, a cut-off of .90 might be more appropriate. Also, the selection of the best cut-off score of the IOP-29 should always be based on the purposes of the examination and on the costs and benefits associated with false positive and false negative classifications. Thus, if the IOP-29 is used as a screening tool, lower probability cut-offs might be preferable so as to maximize NPP.

Despite the IOP-29 consisting of only 29 items, neither the MMPI, the TOMM, nor the PAI outperformed the IOP-29 in discriminating between bona fide patients and simulators. These findings, again, support the potential utility and validity of the IOP-29 as a brief, feigning measure. To further investigate the comparative validity of the IOP-29, future studies should also compare its performance to that of other brief instruments, such as the

SIMS or the M Test. Most importantly, future research should focus on ecologically valid field studies rather than simulation studies.

Despite these encouraging findings, some limitations persist. The most obvious limitation is that we investigated large archival files but did not test the instrument in its final 29-item format, and the great majority were administered in a computerized versus paper-and-pencil format. In other words, in all cases patients and simulators were administered a large number of items that included but were not limited to the final 29 IOP-29 items. Although other instruments have been developed in a similar manner (for example, see Peters, Sunderland, Andrews, Rapee, & Mattick, 2012), this may have affected our results, to some extent. To address this limitation, new data are currently being collected, in various countries and cultural contexts.

Another limitation of these studies, which is actually shared by all simulation studies, is that it is impossible to know whether the behavior of the simulators reflect that of true malingerers in actual real-life situations. To maximize generalizability we took several methodological cautions, such as providing contexts to facilitate malingering, offering extra incentives for the best feigners, cautioning to not exaggerate so as to produce a believable profile, etc. (Rogers, 1997, 2008; Rogers & Bender, 2013; Rogers & Gillard, 2011; Viglione et al., 2001). Nonetheless, it is still unknown whether true malingerers in real-life situations would have obtained the same scores as those produced by our simulator samples. Finally, future research might consider investigating samples with base rates different from .50, such that different cut-off scores could be investigated.

References

Abramsky, A. B. (2005). *Assessment of test behaviors as a unique construct in the evaluation of malingered depression on the inventory of problems: Do test behaviors add significant variance beyond problem endorsement strategies?*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Arbisi, P. A., & Ben-Porath, Y. S. (1995). An MMPI–2 infrequent response scale for use with psychopathological populations: The Infrequency-Psychopathology scale, F(p). *Psychological Assessment, 7*, 424–431.

Archer, R. P., Buffington-Vollum, J. K., Stredny, R.V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84–94.

Beaber, R. J., Marston, A., Michelli, J., & Mills, M. J. (1985). A brief test for measuring malingering in schizophrenic individuals. *American Journal of Psychiatry, 142*, 1478-81.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2 Restructured Form): Manual for administration, scoring, and interpretation. Minneapolis*. University of Minnesota Press.

Blanchard, D.D., McGrath, R.E., Pogge, D. L., & Khadivi, A. (2003). A comparison of the PAI and MMPI-2 as predictors of faking bad in college students. *Journal of Personality Assessment, 80*, 197-205.

Chafetz, M., & Underhill, J. (2013). Estimated costs of malingered disability. *Archives of Clinical Neuropsychology, 28(7)*, 633–639.

Connell, K. (2004). *Detecting simulated versus genuine posttraumatic stress disorder*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Edens, J. F., Poythress, N. G., & Watkins-Clay, M. M. (2007). Detection of malingering in psychiatric unit and general population prison inmates: A comparison of the PAI, SIMS, and SIRS. *Journal of Personality Assessment, 88(1)*, 33-42.

Elhai, J. D., Gold, P. B., Frueh, B. C., & Gold, S. N. (2000). Cross-validation of the MMPI-2 in detecting malingered posttraumatic stress disorder. *Journal of Personality Assessment, 75*, 449-463.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1995). Structured Clinical Interview for Axis I DSM-IV. Disorders - Patient Edition (SCID-I/P). New York: Biometrics Research Department, NY State Psychiatric Institute.

Green, L. (1999). *Computerized versus standard administration of the Inventory of Problems: An examination of reliability and validity*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Green, P., Allen, L. M., & Astner, K. (1996). *The Word Memory Test: A user's guide to the oral and computer administered forms*. Durham, NC: CogniSyst Inc.

Green, R.L. (1991). *The MMPI-2/MMPI: An interpretative manual*. Boston, MA: Allyn & Bacon.

Hawes, S., & Bocaccini, M. (2009). Detection of overreporting of psychopathology on the Personality Assessment Inventory: A meta-analytic review. *Psychological Assessment, 21(1),* 112-124.

Jansak, D. (1996). *The Rorschach comprehensive system depression index, depression heterogeneity, and the role of self-schema*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A study of experts. *Professional Psychology: Research and Practice, 34*, 491–498.

Landis, P. E. (1996). *Detection of simulated posttraumatic stress disorder: A validation study of the Inventory of Problems*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Larrabee, G.J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist, 17*, 54–68.

McCullaugh, J. M. (2011). *The convergent and ecological validity of the Inventory of Problems with a community-supervised, forensic sample*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

McDougall, A. (1996). *Rorschach indicators of simulated schizophrenia*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194-216.

Mellin, D. (1996). *Test-dependent malingering: can the testing situation induce and measure behaviors suggestive of malingering?*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172-175.

Miller, H. A. (2001). *M-FAST: Miller-forensic assessment of symptoms test professional manual*. Odessa, FL: Psychological Assessment Resources.

Morey, L. (1991). *Personality Assessment Inventory: Professional manual.* Tampa, FL:

Morey, L. (1996). *An interpretive guide to the Personality Assessment Inventory (PAI)*. Odessa, FL: Psychological Assessment Resources.

Morey, L. C. (2007). *Personality Assessment Inventory (PAI). Professional Manual (2nd ed.)*. Odessa, FL: Psychological Assessment Resources.

O'Brien, S. M. (2004). *An investigation into the incremental value of test-dependent malingering of schizophrenia*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Peters, L., Sunderland, M., Andrews, G., Rapee, R. M., & Mattick, R. P. (2012). Development of a short form Social Interaction Anxiety (SIAS) and Social Phobia Scale (SPS) using nonparametric item response theory: the SIAS-6 and the SPS-6. *Psychological Assessment, 24(1)*, 66-76.

Pizitz, T. D. (2001). *Detection of malingered mild head injury using the tripartite conceptual model of malingering and the inventory of problems*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego. Psychological Assessment Resources.

Rees, L.M., Tombaugh, T.N., Gansler, D.A., & Moczynski, N.P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment, 10,* 10-20.

Rogers, R. & Bender, S.,D. (2013). Evaluation of malingering and related response styles. In I. B. Weiner (Ed.-in-Chief), J. R. Graham & J. A. Naglieri (Vol. Eds.), *Comprehensive Handbook of Psychology: Assessment Psychology* (2nd Edition Vol. 10, pp. 517-540). Hoboken, NJ: John Wiley & Sons.

Rogers, R. & Salekin, R. (1998). Research report beguiled by Bayes: A re-analysis of Mossman and Hart's estimates of malingering. *Behavioral Sciences and the Law, 16*, 147-153.

Rogers, R. (1997). *Clinical Assessment of malingering and deception*. New York: The Guilford Press.

Rogers, R. (2008). Detection strategies for malingering and defensiveness. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 14–35). New York, NY: Guilford Press.

Rogers, R. (Ed.). (1988). *Clinical assessment of malingering and deception (1st ed.)*. New York: Guilford Press.

Rogers, R., & Gillard, N. D. (2011). Research methods for the assessment of malingering. In B. Rosenfeld & S. Penrod (Eds.), *Research methods in forensic psychology* (pp. 174-188). Hoboken, NJ: Wiley.

Rogers, R., Gillis, J. R., Dickens, S. E., & Bagby, R. M. (1991). Standardised assessment of malingering: Validation of the structured interview of reported symptoms. *Psychological Assessment, 4*, 89–96.

Rogers, R., Gillis, J., Bagby, R., & Monteiro, E. (1991). Detection of malingering on the Structured Inverview of Reported Symptoms (SIRS): A study of coached and uncoached simulators. *Psychological Assessment, 3*, 673-677.

Rogers, R., Harrell, E. H., & Liff, C. D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review, 13*, 255-274.

Rogers, R., Salekin, R. T., Sewell, K. W., Goldstein, A. M., & Leonard, K. (1998). A comparison of forensic and nonforensic malingerers: A prototypical analysis of explanatory models. *Law and Human Behavior, 22*, 353-367.

Rogers, R., Sewell, K. W., & Gillard, N. D. (2010). *Structured Interview of Reported Symptoms (SIRS) and professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Rogers, R., Sewell, K. W., Morey, L. C., & Ustad, K. L. (1996). Detection of feigned mental disorders on the Personality Assessment Inventory: A discriminant analysis. *Journal of Personality Assessment, 67*, 629–640.

Rogers, R., Sewell, K. W., & Ustad, K. L. (1995). Feigning among chronic outpatients on the MMPI-2: An analogue study. *Assessment, 2*, 81–89.

Schaich, D. (2000). *A synthesis of the conceptual and inferential procedures involved with the assessment of malingering: An applied study*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Sellbom, M., Bagby, R. M., & Rogers, R. (2008). *Response styles on multiscale inventories. Clinical assessment of malingering and deception (3rd ed.)*. (pp. 182-206). New York, NY US: Guilford Press.

Slick, D., Hopp, G., Strauss, E., & Thompson, G. B. (2005). *VSVT Victoria Symptom Validity Test*. Odessa, FL: Psychological Assessment Resources.

Smith, G. P., & Burger, G. K. (1997). Detection of malingering: Validation of the Structured Inventory of Malingered Symptomatology (SIMS). *Journal of American Academic Psychiatry and the Law, 25*, 183–189.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*, 217-222

Systems.

Tombaugh, T.N. (1996). *Test of Memory Malingering.* Toronto, Canada: Multi-Health.

Viglione, D. J. & Landis, P. (1994). *The Development of an Objective Test for Malingering*. Paper presented at the biennial meeting American Psychology-Law Society, Santa Fe, New Mexico.

Viglione, D. J., Wright, D., Dizon, N. T., Moynihan, J. E., DuPuis, S., & Pizitz, T. D. (2001). Evading detection on the MMPI-2: Does caution produce more realistic patterns of responding?. *Assessment, 8(3)*, 237-250.

Watson, C. G., Juba, M. P., Manifold, V., Kucala, T., & Anderson, P. E. D. (1991). The PTSD interview: Rationale, description, reliability, and concurrent validity of a DSM-III-based technique. *Journal of Clinical Psycho1ogy, 47*, 179-188.Widows, M. R., & Smith, G. P. (2004). *SIMS: Structured Inventory of Malingered. Symptomatology professional manual*. Odessa, FL: Psychological Assessment.

Weber, K. A. (2008). *Improving External Validity with Simulators in Malingering Research*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Wood, S. (2008). *Unique contributions of performance and self-report methods in the detection of malingered psychotic symptoms*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

DEVELOPMENT OF THE IOP-29

Table 1. Composition of the Combined Sample used for Study 2.

	Patients (n = 166)	Simulators (n = 167)	Total (n = 333)
Diagnosis/Target ($Chi^2_{(3)} < .01$, *n.s.*)			
Head Injury	38	38	76
Psychosis	44	45	89
PTSD	40	40	80
Depression	44	44	88
Gender (*Phi* = .01, *n.s.*)			
Male	74	72	146
Female	92	95	187
Age ($Chi^2_{(4)} = 6.85$, *n.s.*)			
Range	18 to 67	18 to 69	18 to 69
Less than 25	27	37	64
25 to 34	33	47	80
35 to 44	52	40	92
45 to 54	36	28	64
More than 54	18	15	33
Education (*Phi* = .05, *n.s.*)			
Less than College	59	51	110
Some College or Higher	107	116	223
Marital Status (*Phi* = -.10, *n.s.*)			
Married	56	72	128
Other	110	95	205

DEVELOPMENT OF THE IOP-29

Table 2. Diagnostic Efficiency Statistics of the IOP-29 False Disorder Probability Score (Validation Sample, Study 2; *n* = 165).

Cut-off	Sensitivity	Specificity	*Base Rate = .50*			*Base Rate = .25*			*Base Rate = .10*			*Base Rate = .05*		
			PPP	NPP	OCC	PPP	NPP	OCC	PPP	NPP	OCC	PPP	NPP	OCC
0.05	1.00	0.06	0.52	1.00	0.53	0.26	1.00	0.30	0.11	1.00	0.15	0.05	1.00	0.11
0.10	0.95	0.22	0.55	0.82	0.59	0.29	0.93	0.40	0.12	0.98	0.29	0.06	0.99	0.26
0.20	0.94	0.40	0.61	0.87	0.67	0.34	0.95	0.54	0.15	0.98	0.46	0.08	0.99	0.43
0.50	0.81	0.79	0.80	0.80	0.80	0.56	0.93	0.80	0.30	0.97	0.79	0.17	0.99	0.79
0.80	0.47	0.95	0.91	0.64	0.71	0.76	0.84	0.83	0.52	0.94	0.90	0.34	0.97	0.93
0.90	0.34	0.99	0.97	0.60	0.66	0.90	0.82	0.83	0.75	0.93	0.92	0.59	0.97	0.96
0.95	0.13	1.00	1.00	0.53	0.56	1.00	0.78	0.78	1.00	0.91	0.91	1.00	0.96	0.96

PPD = Positive Predictive Power; NPP = Negative Predictive Power; OCC = Overall Correct Classification.

DEVELOPMENT OF THE IOP-29

Table 3. Point Bi-serial Correlations with Group and Area Under the Receiver Operator Characteristic Curves for the IOP-29, MMPI-2, and TOMM.

	Point Bi-serial Correlations with Group[a]	AUC (*SE*)
O'Brien's (2004) Schizophrenic/Psychotic Sub-Sample (n = 88)		
IOP-29 Multiple Weight Score	.61**	.845 (*.042*)
IOP-29 False Disorder Probability Score	.61**	.845 (*.042*)
MMPI-2 Scale F	.44**	.789 (*.051*)
MMPI-2 Scale Fp	.51**	.809 (*.049*)
MMPI-2 Scale Ds-r2	.65**	.880 (*.037*)
Abramsky's (2005) Depression Sub-Sample (n = 85)		
IOP-29 Multiple Weight Score	.67**	.898 (*.037*)
IOP-29 False Disorder Probability Score	.71**	.898 (*.037*)
TOMM-1	-.51**	.760 (*.053*)
TOMM-2	-.58**	.821 (*.048*)
TOMM-tot	-.57**	.814 (*.047*)

[a] Dummy Code: 1 = Simulator; 0 = Bona Fide Patient. *p<.05; **p<.01.

DEVELOPMENT OF THE IOP-29

Table 4. Point bi-serial correlations and AUCs of the IOP-29 and PAI within the Forensic (*n* = 128) and Psychotic (*n* = 90) Samples.

	Point bi-serial Correlation with Group [a]	AUC (*SE*)
Forensic Sample (n = 128)		
IOP-29 – False Disorder Probability Score	.80**	.943 (*.023*)
PAI – Negative Impression Management (T score)	.82**	.966 (*.016*)
PAI – Malingering Index (raw score)	57**	.837 (*.038*)
PAI – Rogers Discriminant Function (raw score)	.68**	.892 (*.030*)
Psychotic Sample (n = 90)		
IOP-29 – False Disorder Probability Score	.70**	.915 (*.028*)
PAI – Negative Impression Management (T score)	.65**	.887 (*.034*)
PAI – Malingering Index (raw score)	.67**	.909 (*.031*)
PAI – Rogers Discriminant Function (raw score)	.66**	.886 (*.037*)

[a] Dummy Code: 1 = Simulator; 0 = Bona Fide Patient. *p<.05; **p<.01.

DEVELOPMENT OF THE IOP-29

Table 5. Hit Rates of the IOP-29 and PAI within the Forensic (n = 128) and Psychotic (n = 90) Samples.

	Group		Diagnostic Efficiency Statistics				
	Patients	Simulators	Sensitivity	Specificity	PPP	NPP	OCC
Forensic Sample (n = 128)							
IOP-29 – False Disorder Probability Score			.72	1.00	1.00	.78	.86
p < .50	64	18					
p ≥ .50	0	46					
PAI – Negative Impression Management			.75	.98	.98	.80	.87
T < 81	63	16					
T ≥ 81	1	48					
PAI – Malingering Index			.08	.98	.83	.52	.53
Raw < 3	63	59					
Raw ≥ 3	1	5					
PAI – Rogers Discriminant Function			.72	.92	.90	.77	.82
Raw < 0	59	18					
Raw ≥ 0	5	46					
Psychotic Sample (n = 90)							
IOP-29 – False Disorder Probability Score			.82	.80	.80	.82	.81
p < .50	36	8					
p ≥ .50	9	37					
PAI – Negative Impression Management			.82	.73	.76	.81	.78
T < 81	33	8					
T ≥ 81	12	37					
PAI – Malingering Index			.89	.76	.78	.87	.82
Raw < 3	34	5					
Raw ≥ 3	11	40					
PAI – Rogers Discriminant Function			.96	.67	.74	.94	.81
Raw < 0	30	2					
Raw ≥ 0	15	43					

PPP = Positive Predictive Power; NPP = Negative Predictive Power; OCC = Overall Correct Classification. Cut scores for the PAI scales were defined based on Hawes and Bocaccini's (2009) meta-analytic review (NIM and MAL) and on Morey's (2007) PAI manual (RDF). PPP, NPP, and OCC are calculated using the base rate of our sample, which is equal to .50.